CENTRAL INSTITUTE OF INDIAN LANGUAGES
DEPARTMENT OF HIGHER EDUCATION
Ministry of Human Resource Development, Government of India
Manasagangotri, Mysore - 570 006

MINUTES OF THE SECOND PROJECT ADVISORY COMMITTEE MEETING OF THE LINGUISTIC DATA CONSORTIUM FOR INDIAN LANGUAGES (LDC-IL) HELD ON JUNE 9, 2008 AT 11.30 a.m.

I. Welcome

Prof. Udaya Narayana Singh, Director, Central Institute of Indian Languages and Chairperson, Linguistic Data Consortium for Indian Languages (LDC-IL) welcomed the Members for the Second Project Advisory Committee meeting. He explained the constraints for not having another one meeting during the past financial year. The Director also took this opportunity to brief the members about the new projects sanctioned by the Ministry, especially about the National Translation Mission, which will have a bearing on the LDC-IL Project.

II. Agenda Items

After the Welcome, the agenda items were taken up in the order.

1. The Minutes of the First Project Advisory Committee Meeting of the Linguistic Data Consortium for Indian Languages (LDC-IL) held on June 5, 2007 were confirmed.

2. The Mission Statement of the LDC-IL namely, the following was approved: "Annotated, quality language data (both-text and speech) and tools in Indian languages to Individuals, Institutions Industry etc., for Research and Development - Created in house, through outsourcing and, acquisition".

3. Dr. B. Mallikarjun, Reader cum Research Officer & Head, LDC-IL made a presentation on Action Taken Report on the recommendations of the First PAC meeting, progress made in the work of LDC-IL from June 5, 2007 to June 8, 2008 and proposed certain targets for the year 2008-09. The proposed targets and the details of progress made are given in **Annexure – 2** in a tabular form.

III. Action Taken Report in respect of Working Groups

(a) The Working Groups on Licensing Issues, Natural Language Processing, Speech and Speech Deficiencies have met on August 3, 2007 at Pune, August 6, 2007 at Hyderabad, and November 29, 2007 at Mysore respectively and deliberated various issues relating to LDC-IL. The Working Group on Licensing Issues is expected to have another meeting to make specific and concrete recommendations. The Natural Language Processing Group has standardized POS Tag set and XML tag set (given in **Annexure - 3**). The group has assigned tasks to members of the group to think and provide write ups on future directions. The group will study the drafts prepared by the individuals in order to prepare a document on future directions.

(b) The Scholars of the Speech Group have met several times and arrived at standards for Speech Data capturing and Annotation. The Speech/Language Development Group has met and some of the personnel had sent their projects to the LDC-IL for grant in aid. However, they have been asked to recast the same.

(c) The Character Recognition Group could not meet due to various reasons. Prof. B.B. Chaudhuri, the Chairman of the Group said that he had informal discussion with members of the group and that they would try to give the scanned texts for LDC-IL. It was also agreed that this Working Group will meet on the sidelines of the next PRSG of

the MCIT Meeting to make specific recommendations regarding the tasks to be undertaken by the LDC-IL in this area.

(d) The following Standards for Language Data were presented, and discussed. They are accepted.

Text Corpora

- Text in UNICODE
- Markup: SGML standard
- POS tagging: Extendable and expandable decided by the NLP Group on August 6, 2007.
-

Speech Corpora

- Rate of sampling - Multiples of 8 kHz. The purpose and the rate of sampling to be uniform.
- Transliteration scheme- LDC-IL standard
- Annotation - PRAAT, Wave surfer
- Pronunciation Dictionary – Format (All placed before the PAC)

(For the convenience of the PAC members absent in the PAC the full text on standards for speech is enclosed with this as **Annexure - 4**).

(e) The Copies of the First versions of the Training Modules prepared by Prof. Dipti Misra et al., on POS Tagging and Chunking, Prof. Amba Kulkarni on Morphological analyzer, Prof. Pushpak Bhattacharya on Sense tagging and Prof. Peri Bhaskararao on Collecting Speech data were given to the Project Advisory Committee. These modules will be used by the LDC-IL for in-house training as well as for the training that it will conduct elsewhere for collecting and annotating language data.

(f) It was noted that the following programmes conducted and sponsored by LDC-IL along with the reports were placed in the PAC meeting:

- Winter School on Speech and Audio Processing (WiSSAP 2008) from 2nd to 5th January 2008 held at IIT, Madras.
- Workshop on 'Advanced Course in Computational Linguistics' held from 16th to 25th March 208 by the Dravidian University, Kuppam.
- Workshop on 'Speech Sciences' held at CIIL, Mysore from 10th - 21st March 2008.
- Selection Workshop/Test conducted to recruit staff for the LDC-IL Project from 17th - 21st March 2008 at CIIL, Mysore.
- LDC-IL staff training May 19 to June 13, 2008.

(g) A list of tasks that were recommended by the First Project Advisory Committee but not conducted by the LDC-IL was also provided to the PAC.

(h) In the absence of specifically appointed staff for the LDC-IL, the institute using the Workshop mode resource persons has created monolingual text corpora, parallel corpora and speech corpora. In doing so only availability of Resource Persons was taken into consideration and language priority was not considered. The statistical details were presented before the Committee as a part of the progress report.

IV. Recommendations

The members deliberated these and made the following recommendations:

(1) The LDC-IL has a national role and has to function its assigned responsibility as a nodal agency. Therefore, the Linguistic Data Consortium for Indian Languages (LDC-IL) could be visualized as a repository of language/linguistic resources and tools for Indian languages. An attempt should be made by the LDC-IL to contact all the NLP groups and institutions and collect them. After collecting they have to be tested and validated. The resources and tools that are up to the standard can be licensed by the LDC-IL under its Licensing Policy.

(2) A Road Map for the work of LDC-IL for the current year as well as for the Eleventh plan period has to be drawn and placed before the next PAC meeting. A priority list of languages has to be prepared for creation of resources.

(3) Regarding all Speech for speech corpus of languages the number of hours has to be 20 hrs. and not 10 hrs.

(4) While preparing and procuring data, end users needs have to be kept in mind. The focus has to be end user.

(5) For evaluation process for each kind of data, tool etc., matrixes have to be evolved. Bench marking, good standards etc., have to be developed.

(6) Indian Sign Language Vocabulary has to be developed.

(7) Dr. Anupam Basu of IIT Kharagpur be co-opted for NLP group.

V. Other Matters

(a) The existing 10 vacant positions will be filled by academic persons and they will be re-designated as Research Assistants (Senior) and Research Assistants (Junior). The existing academic persons under Technical positions will also be re-designated as Research Assistant (Senior).

(b) The member representing IBM Shri Abhijit Dutta said that the Workshop on Speech Recognition they intend to do in the previous financial year shall be conducted in the next few months.

(c) The members present were requested to send proposals for Seminars, events, workshops, training programmes etc.

(d) Working Group on Speech Deficiency will be **renamed as Working Group on Speech Language Development.**

(e) Grant-in-Aid : In case if some grantee does not show adequate progress to commensurate with the release of funds, further release of funds will be stopped and action will be taken as per the agreement.

(f) The next meeting of the **LDC-IL PAC** will be held at Mysore in the **last week of November 2008**. All the members shall keep this in mind.

The meeting ended with thanks to the Chair.

Udaya

(UDAYA NARAYANA SINGH)
Chairperson & Director, Linguistic Data
Consortium for Indian Languages,
Central Institute of Indian Languages, Mysore